Discriminative Adaptive Training and Discriminative Adaptation

Lan Wang

December 5th 2003

Cambridge University Engineering Department

Overview

- Introduction
- Speaker adaptive training
- Discriminative training
- Discriminative speaker adaptive training (DSAT)
  - discriminative linear transform estimation
  - DSAT model (canonical model) parameter re-estimation
- Experiments
  - DSAT experiments
  - discriminative adaptation experiments
- Discussion and conclusion

Introduction

- Data for training state-of-the-art LVCSR systems:
  - specifically collected data.
  - large amount of "found" (conversation & broadcast) data.

- Complex and uncontrolled variabilities in the training data.
  - thousands of speakers,
  - noisy background (environment),
  - diversity of channels.

- Traditional acoustic modeling:

  all the data is pooled together as a single block for training.

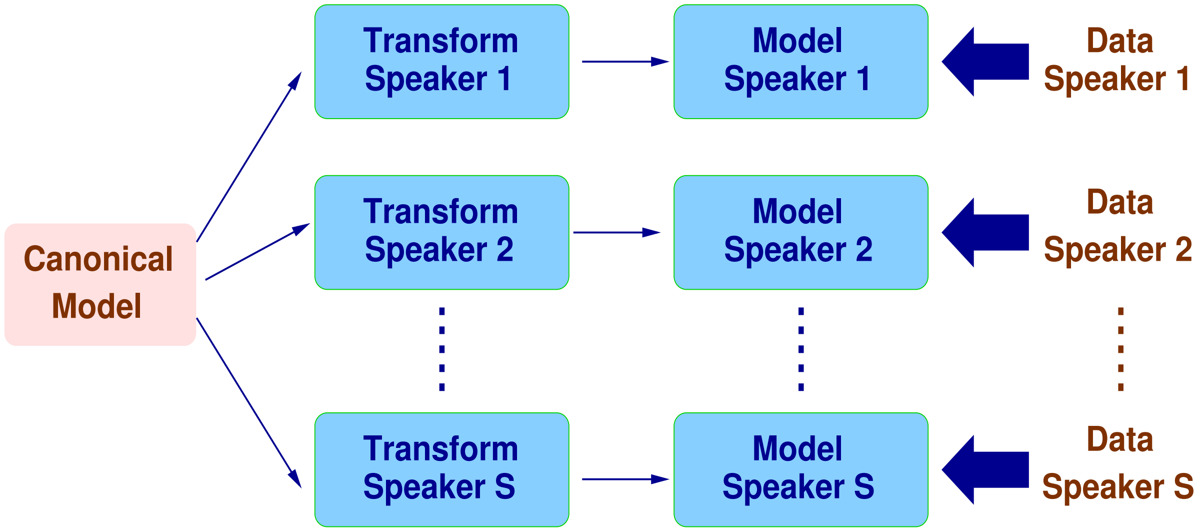
Introduction (II)

- Current acoustic modeling:
  - to remove those variations in the training data,
  - to build speaker (acoustic condition) independent HMM sets.
- Feature-space normalization:
  - cepstral mean normalization (CMN)
  - cepstral variance normalization (CVN).
  - vocal tract length normalization (VTLN).
- Model-space transformation: adaptive training

  linear transforms are extensively used for both training and adaptation.

Speaker Adaptive Training

- SAT: speaker-specific train-set transforms are applied to the HMM parameter optimization so as to construct a canonical HMM set.



- The canonical models with testing adaptation perform better than non-SAT models.

Speaker Adaptive Training (II)

- Maximum likelihood (ML) framework for transform estimation.
  - maximum likelihood linear regression (MLLR):

$$\hat{\mu}_m = \mathbf{A}\mu_m + \mathbf{b} = W\xi_m$$

  - constrained MLLR:
    the same transforms are used to adapt means and covariances

$$\hat{\mathbf{o}}(t) = \mathbf{A}\mathbf{o}(t) + \mathbf{b} = W\zeta(t)$$

- Canonical model parameter re-estimation under ML criterion.

Discriminative Training

- Maximum mutual information (MMI) criterion.

$$\mathcal{F}_{MMI}(\lambda) = \sum_{r=1}^{R} \log \frac{P_\lambda(\mathcal{O}_r|\mathcal{M}^{w_r})^\kappa P(w_r)^\kappa}{\sum_{\hat{w}} P_\lambda(\mathcal{O}_r|\mathcal{M}^{\hat{w}})^\kappa P(\hat{w})^\kappa}$$

- Minimum phone error (MPE) criterion.

$$\mathcal{F}_{MPE}(\lambda) = \sum_{r=1}^{R} \frac{\sum_s p_\lambda(\mathcal{O}_r|\mathcal{M}^{w_s})^\kappa P(w_s)^\kappa RawAccuracy(w_s, w_r)}{\sum_u p_\lambda(\mathcal{O}_r|\mathcal{M}^{w_u})^\kappa P(w_u)^\kappa},$$

where $RawAccuracy(w_s, w_r)$ measures the accuracy of hypothesis w_s.

- Lattice-based framework.

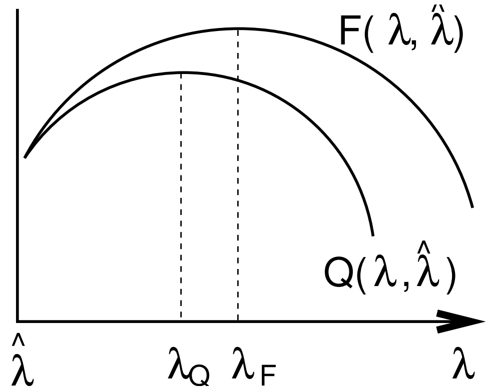
Optimization of Discriminative Criteria

- Strong-sense auxiliary function for ML optimization.

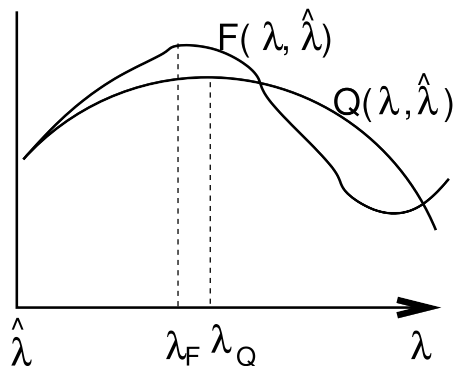
$$\mathcal{Q}(\lambda,\hat{\lambda})-\mathcal{Q}(\hat{\lambda},\hat{\lambda})\leq\mathcal{F}(\lambda)-\mathcal{F}(\hat{\lambda})$$

- Weak-sense auxiliary function for the optimization of discriminative criteria.

$$\frac{\partial}{\partial\hat{\lambda}}\mathcal{Q}(\lambda,\hat{\lambda})\Big|_{\hat{\lambda}=\lambda}=\frac{\partial}{\partial\hat{\lambda}}\mathcal{F}(\hat{\lambda})\Big|_{\hat{\lambda}=\lambda}$$



(a)Strong Sense



(b) Weak Sense

Discriminative SAT

- Discriminative SAT (DSAT):

  discriminative criteria are used consistently to construct the canonical models in two sequential steps:

  – discriminative linear transform (DLT) generation,
  – discriminaitve model (canonical model) parameter re-estimation.

- DSAT implementations:

  – DLT & canonical model re-estimation.
  – constrained DLT & canonical model re-estimation.

- A simplified implementation:

  ML-based transform is used for discriminative model parameter re-estimation.

DLT Estimation

- Applying linear transforms to tune Gaussian components.
  - mean transform W.
  - diagonal (full) variance transform: $\hat{\Sigma}_m = \mathbf{H}^T \Sigma_m \mathbf{H}$
- Supervised and unsupervised adaptation.
- Weak-sense auxiliary function for the optimization MMI/MPE-based DLT.

$$\mathcal{Q}_{MMI}(W, \hat{W}) = \mathcal{Q}^{num}(W, \hat{W}) - \mathcal{Q}^{den}(W, \hat{W}) + \mathcal{Q}_{sm}(W, \hat{W})$$

$$\mathcal{Q}^{num}(W, \hat{W}) = \sum_r \sum_m \sum_t \gamma_m^{num}(t) \log \mathcal{N}(\mathbf{o}(t), \hat{W}\xi_m, \Sigma_m)$$

$$\mathcal{Q}_{sm}(W, \hat{W}) = \sum_r \sum_m D_m \left[-\frac{1}{2} \left(\log |\hat{\Sigma}_m| + (W\xi_m - \hat{W}\xi_m)^T \hat{\Sigma}_m^{-1} (W\xi_m - \hat{W}\xi_m) + \Sigma_m \hat{\Sigma}_m^{-1} \right) \right]$$

DLT Estimation (II)

- Transform estimation for each row: $\mathbf{w}^{(i)} = \mathbf{G}^{(i)^{-1}}\mathbf{k}^{(i)}$
  - ML accumulator:

$$\mathbf{G}^{(i)} = \sum_m \frac{1}{\sigma^2_{m(i)}} \gamma_m \xi_m \xi_m^T$$

  - MMI/MPE accumulator:

$$\mathbf{G}^{(i)} = \sum_m \frac{1}{\sigma^2_{m(i)}} \Big((\gamma_m^{num} - \gamma_m^{den}) + D_m \Big) \xi_m \xi_m^T$$

    where $D_m = E\gamma_m^{den}$ with constant E.

- The num/den occupancies are computed as in MMI/MPE training.
- I-smoothing for MPE-based DLT: using ML statistics.

DLT for DSAT Model Re-estimation

- Optimize MMI/MPE objective functions by applying transforms to the models.
- DLT for MMI/MPE-SAT model parameter re-estimation (e.g. for means).

$$\hat{\mu}_m = \mathbf{M}_m^{-1}\mathbf{V}_m + \mu_m$$

  – ML statistics:

$$\mathbf{M}_m = \sum_{s}^{S} \gamma_m \; {\mathbf{A}^{(s)}}^T \; \Sigma_m^{-1} \; \mathbf{A}^{(s)}$$

  – MMI/MPE statistics:

$$\mathbf{M}_m = \sum_{s=1}^{S} \Big((\gamma_m^{num} - \gamma_m^{den}) + D_m \Big) \; {\mathbf{A}^{(s)}}^T \; \Sigma_m^{-1} \; \mathbf{A}^{(s)}$$

- Computational expensive.

Constrained DLT Estimation

- Weak-sense auxiliary function for the optimization of MMI/MPE-based constrained DLT.

- An iterative optimization to estimate transforms, like constrained MLLR.
  - ML accumulator:
$$\mathbf{G}^{(i)} = \sum_m \frac{1}{\sigma^2_{m(i)}} \sum_t \gamma_m(t)\zeta(t)\zeta(t)^T$$
  - MMI/MPE accumulator:
$$\mathbf{G}^{(i)} = \sum_m \frac{1}{\sigma^2_{m(i)}} \left(\sum_t \left(\gamma_m^{num}(t) - \gamma_m^{den}(t) \right) \zeta(t)\zeta(t)^T + D_m \begin{bmatrix} 1 & \tilde{\mu}_m^T \\ \tilde{\mu}_m & \tilde{\Sigma}_m + \tilde{\mu}_m \tilde{\mu}_m^T \end{bmatrix} \right)$$

- The num/den occupancies are computed as in MMI/MPE training.

- Baseclass I-smoothing technique for MPE-based constrained DLT.

Constrained DLT for DSAT Model Re-estimation

- Constrained DLT for MMI/MPE-SAT model parameter re-estimation.

- Applying to the features makes canonical model parameter re-estimation more straightforward.

- The same updating formulas for MMI/MPE training can be used with adapted observations.

$$\begin{aligned}\hat{\mu}_m &= \frac{\theta_m^{num}(\hat{\mathcal{O}}) - \theta_m^{den}(\hat{\mathcal{O}}) + D_m\mu_m}{\{\gamma_m^{num} - \gamma_m^{den}\} + D_m} \\ \hat{\sigma}_m^2 &= \frac{\theta_m^{num}(\hat{\mathcal{O}}^2) - \theta_m^{den}(\hat{\mathcal{O}}^2) + D_m(\sigma_m^2 + \mu_m^2)}{\{\gamma_m^{num} - \gamma_m^{den}\} + D_m} - \hat{\mu}_m^2\end{aligned}$$

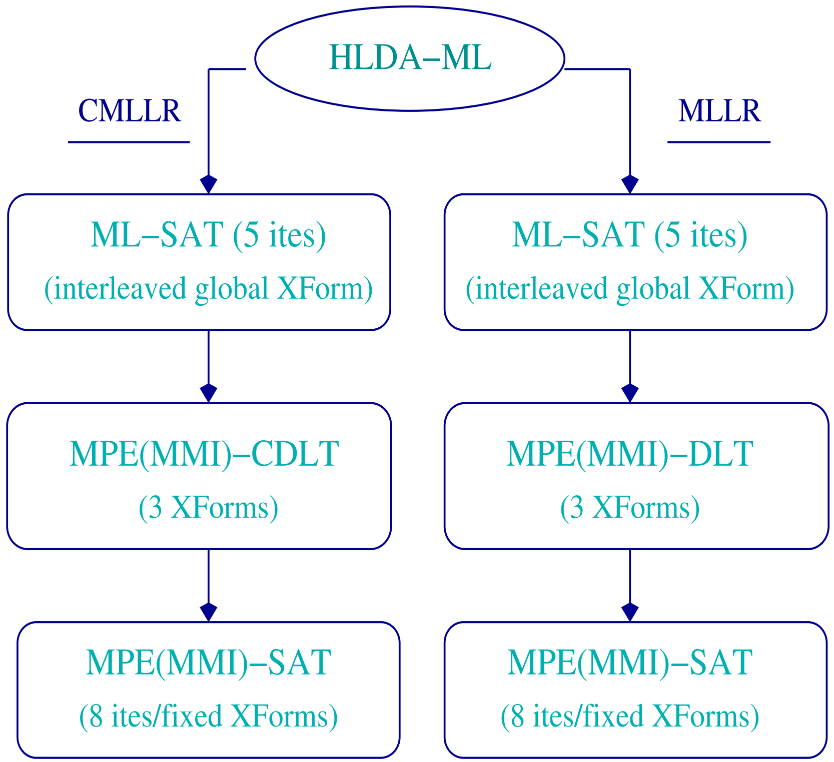
- The num/den statistics are calculated in the same way as MMI/MPE training.

Experiments setup

- Experiments on conversational telephone speech (CTS) transcription.
  - Training set: 76 hours CTS data/1118 conversation sides.
  - Test set ($dev01$): 6 hours CTS data/118 conversation sides.
- The front-end:
  - MF-PLP cepstral parameter $(+\Delta, +\Delta\Delta + \Delta\Delta\Delta)$,
  - HLDA projection (52 dim to 39 dim),
  - VTLN analysis.
- Basic GI HMM sets: 5920 tied-states/12 Gaussian components.

Training setup

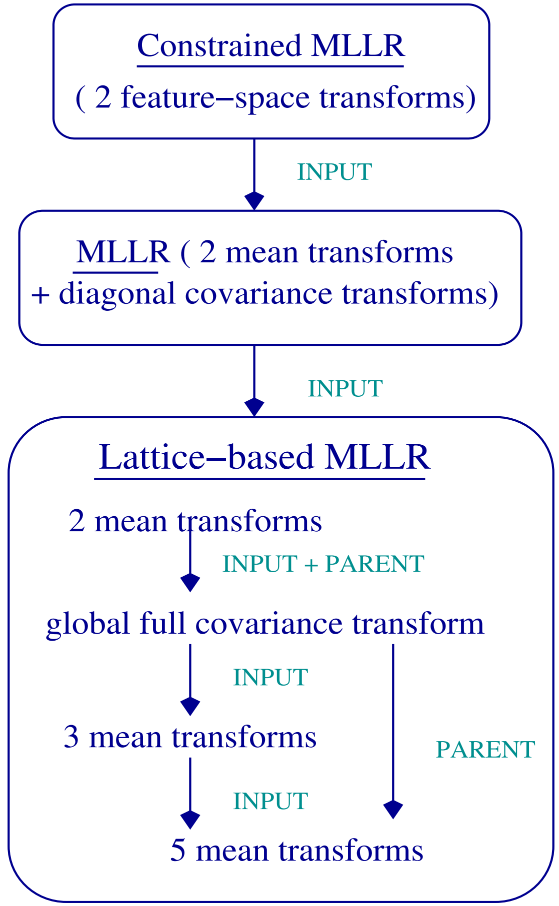
- Starting with the HLDA-ML model.
- Five iterations of interleaved (global) transform estimation and model parameter updating for ML-SAT models.
- Generating unconstrained/constrained MMI/MPE-based DLT (3 transforms per side).
- Eight iterations model parameter updating with fixed transforms for MMI/MPE-SAT models.



Testing setup

- Unsupervised style testing adaptation in a sequential process.
  - 1-best MLLR: transcription assumed correct.
  - Lattice MLLR: used a recognition lattice for forward-backward pass rather than a single model sequence.
- Lattice rescoring to evaluate the discriminative SAT models.

  (lattice generation in the same way as CTS RT03 system)



DSAT with Constrained Linear Transform

	transform generation/parameter re-estimation
MMI-SAT(+CMLLR)	constrained MLLR/MMI
MMI-SAT(+MMI_CDLT)	MMI-based constrained DLT/MMI
MPE-SAT(+CMLLR)	constrained MLLR / MPE
MPE-SAT(+MMI_CDLT)	MMI-based constrained DLT/MPE
MPE-SAT(+MPE_CDLT)	MPE-based constrained DLT/MPE

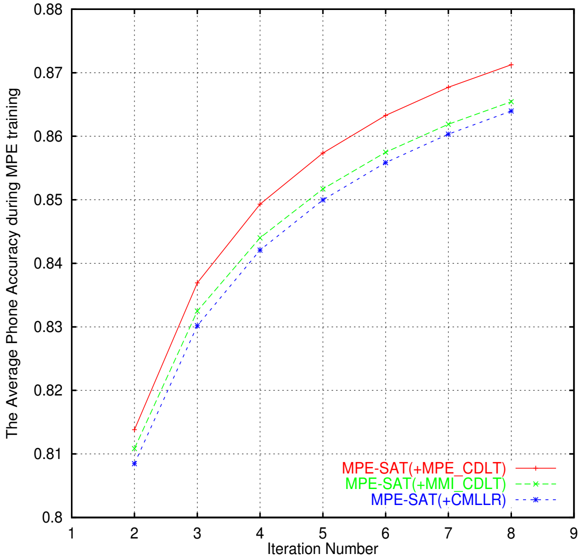
DSAT systems with constrained MLLR/DLT.

Systems	SW-I	SW-II	Cell	total
MMI	21.1	33.4	33.1	29.2
MMI-SAT(+MMI_CDLT)	20.3	32.9	32.6	28.6
MPE	20.2	33.0	32.7	28.6
MPE-SAT(+MPE_CDLT)	20.1	31.8	31.8	27.8

%WER on test set $dev01$ after (1-best) constrained MLLR adaptation

- MMI/MPE-SAT give 0.6%-0.8% abs lower WER than non-SAT MMI/MPE systems.

MPE-SAT with Constrained Linear Transform



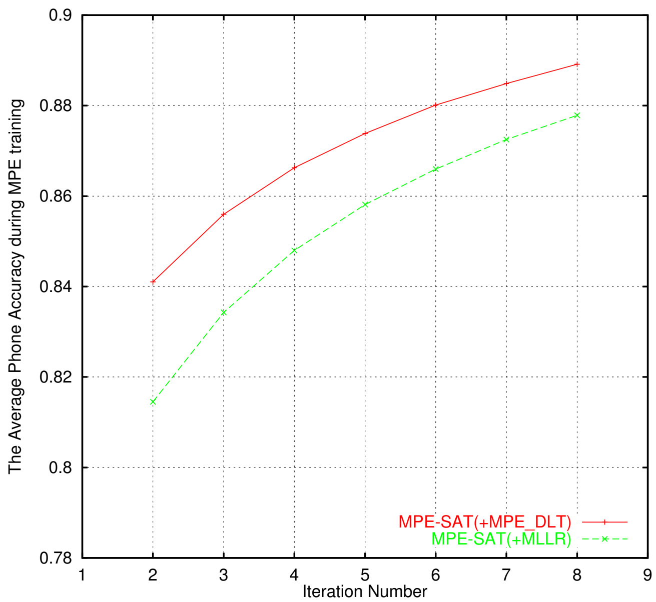
Average phone accuracy during MPE-SAT training.

Systems	lattice MLLR
MPE	27.9
MPE-SAT(+CMLLR)	27.0
MPE-SAT(+MMI_CDLT)	26.9
MPE-SAT(+MPE_CDLT)	26.9

%WER for MPE-SAT systems on $dev01$ after lattice-based MLLR adaptation.

- During training, MPE-SAT with MPE_CDLT outperforms the simplified implementation.
- MPE-SAT with MPE_CDLT just improves the WER by abs 0.1% over MPE-SAT with CMLLR.

MPE-SAT with Unconstrained Linear Transform



Average phone accuracy during MPE-SAT training

	transform generation/ parameter re-estimation
MPE-SAT(+MLLR)	MLLR / MPE
MPE-SAT(+MPE_DLT)	MPE-based DLT/MPE

MPE-SAT systems with MLLR/DLT.

Systems	lattice MLLR
MPE-SAT(+MLLR)	27.0
MPE-SAT(+MPE_DLT)	27.0

%WER for MPE-SAT on $dev01$ after lattice-based MLLR adaptation.

- During training, MPE-SAT with MPE_DLT significantly outperforms the simplified implementation.
- MPE-SAT with MPE_DLT gets almost same performance as MPE-SAT with MLLR.

DLT for Supervised Adaptation

- Supervised adaptation on WSJ task.
- The front-end: 39 dimensional MF-PLP features.
- The cross-word triphone HMMs.
  - ML training.
  - 6399 states/12 Gaussians.
- Testing set: NAB Spoke 3 (s3-dev/s3-eval) with enrollment set.
- H-criterion DLT (a version of MMI criterion) and MPE-based DLT:
  - mean + diagonal variance transforms.
  - regression tree with 16 baseclasses for sp/1 baseclass for sil.

DLT for Supervised Adaptation (II)

Test sets	iterations	MLLR	H-cri	MPE-DLT
s3-dev	1 ite	13.2	12.4	12.2
s3-eval	1 ite	11.1	10.3	10.1
s3-dev	3 ite	12.4	11.9	11.8
s3-eval	3 ite	10.4	10.1	10.0

%WER on NAB Spoke 3 after MLLR, H-criterion and MPE-based DLT adaptation.

- MPE-based DLT achieves 1% abs WER reduction over MLLR and 0.2% over H-criterion DLT (after 1 iteration).

- MPE-based DLT converges fast.

MPE-based DLT for Unsupervised Adaptation

- The cross-word triphone HMMs built with MPE training on CTS transcription.
- MLLR and MPE-DLT: 2 mean+diagonal variance transforms
- Using the hypothesis as supervision:
  - 1-best Viterbi outputs after lattice MLLR adaptation and confusion network (CN) decoding (WER: 27.0%).

Adaptation	hypothesis		true trans
MLLR	27.7	(+CN) 27.0	26.1
MPE-DLT	27.3	(+CN) 26.9	23.2

%WER on *dev01sub* for MPE system.

- For unsupervised style, MPE-based DLT gets 0.1% gain after CN decoding over MLLR.

Discussions and Conclusions

- MMI/MPE-SAT can improve the performance by 0.7%-1.0% compared with non-SAT MMI/MPE training.

- Using MLLR/CMLLR to build MMI/MPE-SAT models is a simplified implementation.

- Using consistent discriminative criteria for MMI/MPE-SAT can give slight improvements under current testing adaptation scheme.

- DLT to adapt discriminative SAT models with unsupervised style.
  - Confidence score to accumulate more confident statistics.
  - Numerator lattices.